DEEP DOMAIN ADAPTATION BY WEIGHTED ENTROPY MINIMIZATION FOR THE CLASSIFICATION OF AERIAL IMAGES

D. Wittich

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany - wittich@ipi.uni-hannover.de

Commission II, WG II/6

KEY WORDS: Domain Adaptation, Aerial Images, Classification, Fully Convolutional Networks, Entropy minimization

ABSTRACT:

Fully convolutional neural networks (FCN) are successfully used for the automated pixel-wise classification of aerial images and possibly additional data. However, they require many labelled training samples to perform well. One approach addressing this issue is semi-supervised domain adaptation (SSDA). Here, labelled training samples from a source domain and unlabelled samples from a target domain are used jointly to obtain a target domain classifier, without requiring any labelled samples from the target domain. In this paper, a two-step approach for SSDA is proposed. The first step corresponds to a supervised training on the source domain, making use of strong data augmentation to increase the initial performance on the target domain. Secondly, the model is adapted by entropy minimization using a novel weighting strategy. The approach is evaluated on the basis of five domains, corresponding to five cities. Several training variants and adaptation scenarios are tested, indicating that proper data augmentation can already improve the initial target domain performance significantly resulting in an average overall accuracy of 77.5%. The weighted entropy minimization improves the overall accuracy on the target domains in 19 out of 20 scenarios on average by 1.8%. In all experiments a novel FCN architecture is used that yields results comparable to those of the best-performing models on the ISPRS labelling challenge while having an order of magnitude fewer parameters than commonly used FCNs.

1. INTRODUCTION

The automated pixel-wise classification of multispectral aerial orthophotos (MSI) and possibly additional data like digital surface models (DSM) is a highly relevant task e.g. for the automated generation or updating of maps. One way to address this task is based on machine learning techniques, where labelled training samples are used to train a classification model. Currently, the best performance in a wide range of applications is achieved by deep neural networks, in particular by variants of fully convolutional neural networks (FCN) (Long et al., 2015a). FCNs are highly scalable classification models that can learn very complex mappings between input and output, if enough training data, representative for the classification task, is available. If this requirement is not fulfilled, the trained model is very likely to overfit to the training data and, thus, to perform badly on unseen data. However, creating more training data usually implies manual labelling, a costly and time-consuming task that should be avoided if possible. To that end, much research is carried out to prevent overfitting without the requirement of additional data by either regularizing the model properly or by artificially increasing the variety of the training data, e.g. by data augmentation (Shorten and Khoshgoftaar, 2019).

Another strategy, referred to as *transfer learning* (TL) (Pan & Yang, 2010), is to transfer knowledge from a source domain, in which training samples are abundant, to a target domain, where only a limited amount or no training data are available. *Domain Adaptation* (DA) is a specific setting of TL, where the domains are assumed to differ only by the joint distribution of the features and the class labels. Regarding the task of aerial image classification (AIC), this corresponds e.g. to a situation where labelled images from one city (source domain) should be used to classify images of another city (target domain) taken with the same sensor type and considering the same class structure. However, the objects in the target domain may have a different appearance, thus, a classifier that was trained only on the source domain possibly performs badly on the target domain.

In the present work the setting of DA is addressed where *only unlabelled samples* from the target domain are used to adapt a classifier from the source to the target domain. According to Tuia et al. (2016) this setting is referred to as *semi-supervised DA*. This setting is particularly interesting, since unlabelled samples from the target domain are always available because they are to be classified in the first place. However, SSDA is known to be very challenging and can even result in a *negative transfer*, denoting a decreased performance on the target domain after adaptation compared to training on the source domain only. Analogously, an improvement is called *positive transfer*. SSDA is highly relevant when it comes to the classification of aerial images because on the one hand there is only a very limited amount of freely available data with annotations (Zhu et al., 2017) and on the other hand the appearance of both natural and man-made objects in aerial images has a huge variability, making it difficult for a model to perform well across different domains. These factors can lead to huge *domain-gaps* in AIC. Although this is certainly a huge challenge, recent advantages in methods for SSDA in related domains like street scene segmentation indicate that these methods can compensate existing domain-gaps in AIC and, thus, increase the applicability of neural networks. However, there is hardly any work addressing SSDA for AIC with deep neural networks and no work was found that considers imbalanced class distributions.

In this paper, a two-step strategy for SSDA for the pixel-wise classification of aerial images is proposed. First, a model is trained in a supervised way on the source domain, referred to as source training. In the second step, the model is adapted to a target domain by applying implicit instance transfer. Following Vu et al. (2019) this is realized by minimizing the mean entropy

of the pixel-wise target domain class predictions. However, the direct entropy minimization is assumed to perform badly in cases where the classes are highly unbalanced, as often in AIC. To that end, a novel weighting technique is introduced. To increase the stability of the adaptation, pixels that are close to a predicted object boundary are not considered in the instance transfer.

In order to evaluate the proposed method, FCNs are trained on five different (source) domains and adapted to the respective four other (target) domains, resulting in 20 adaptation scenarios. By varying the amount of data augmentation during source domain training, the respective influence on the initial target domain performance but also on the adaptability are investigated. In particular, it is assumed that strong data augmentation during source training increases the initial performance on the target domain. The method is compared to the regularized, direct entropy minimization, as proposed in (Vu et al., 2019), to validate the benefit of the proposed variant. A novel FCN architecture is used that has much fewer parameters than common FCNs without a considerable loss of performance. The architecture is mainly based on the combination of partial padding (Liu et al., 2018) and dilated convolutions (Fisher and Vladlen, 2016), assembled in residual layers (Szegedy et al., 2017). The performance of the architecture, it is evaluated on the *Vaihingen* benchmark.

The scientific contributions of this paper are as follows:

- A two-step approach for SSDA based on entropy minimization is proposed and applied to the task of aerial image classification with neural networks. A pixel-wise weighting strategy based on the statistics of semi-labels and predicted object boundaries is proposed that improves both the success rate and the performance of the adaptation.
- As an additional, minor contribution, an architecture for a fully convolutional neural network is proposed. By combining residual layers with partial padding and dilated convolutions the network achieves a performance close to the state of the art while requiring much fewer parameters.
- Lastly, the influence of data augmentation on the initial domain gap and on a succeeding domain adaptation is investigated. It is shown that by using proper data augmentation the domain-gap can be reduced significantly.

2. RELATED WORK

In this section, the state of the art in SSDA in computer vision and photogrammetry is discussed, focussing on the task of pixel-wise classification using FCNs. According to Tuia et al. (2016) SSDA can be either based on representation transfer or on instance transfer. In the following, the two approaches are discussed in further detail.

Representation transfer tries to find mappings from the feature spaces of both domains to a common representation space such that a shared classifier can be applied. In remote sensing, this is often done by finding a mapping that minimizes a statistical distance between the domains, e.g. the maximum-mean discrepancy (MMD) (Matasci et al., 2015). This approach was transferred to neural networks for the task of assigning a single class label to an image in (Long et al., 2015b). Ganin et al. (2015) introduced the concept of domain adversarial training and showed that this approach is superior to minimizing the MMD. The concept of domain adversarial training was also frequently used for the pixel-wise classification with FCNs, e.g. in (Huang et al., 2018), (Hoffmann et al., 2018) and (Zhang et al., 2018) for the semantic segmentation of street scenes. While Zhang et al. (2018) apply the domain discriminator to the final layer of the classification network, Huang et al. (2018) propose to perform the representation transfer in multiple layers of the network. Hoffmann et al. (2018) apply the representation transfer to one intermediate layer of the network. Although all above mentioned approaches yield stable improvements, they are tailored to street scene classification. In (Wittich and Rottensteiner, 2019) the concept of domain adversarial training was applied to the task of aerial image classification. The authors achieve a stable positive transfer of around 1-5% in overall accuracy, evaluated on three domains with three classes. They show that domain adversarial training is highly susceptible to large differences in the marginal class distribution of source and target domain. As the performance of this method highly depends on the network architecture and the adaptation setup seems to be difficult to tune, in this work the alternative concept of instance transfer is explored.

An alternative way of representation transfer is related to the input space of the model. Several approaches for street scene classification (Hoffmann et al., 2018), (Zhang et al., 2018) use image-to-image translation techniques in order to create versions of target domain images that keep their semantic information but look like being drawn from the source domain. To classify a target sample, it is processed by the image-to-image translation network before being fed to the source domain classifier. Although this approach improves the results in street scene segmentation, preliminary experiments on using it for AIC have not shown a significant improvement. A related approach was successfully applied by Tasar et al. (2019) and Benjdira et al. (2019) to AIC. Both papers propose to learn an image-to-image translation network from source to target domain. The domain-gap is reduced by fine-tuning a network initially trained on labelled source domain samples with translated source domain images while keeping the original label maps. Despite the success of these methods, they are not proven to outperform classical augmentation techniques like random modification of brightness and contrast. In this paper augmentation techniques based on strong radiometric and geometric modifications are used and compared to a weak geometrical augmentation to investigate how much this affects the initial domain-gap.

Instance transfer aims at adapting the classifier from the source to the target domain by using semi-labelled samples, i.e. target samples receiving their class labels from the current state of the classifier, e.g. (Bruzzone et al., 2008). Approaches based on instance transfer represent the second largest research branch of SSDA for the task of pixel-wise classification. Addressing the task of street scene segmentation, Zou et al. (2018) propose a class-balanced self-training, where they jointly train a network on labelled source domain data and target domain samples with semi-labels. For each class, they select the semi-labelled samples with the respectively highest confidence. Such a class-balancing is shown to be necessary, when dealing with imbalanced class-distributions. Based on the source domain samples, they further compute a spatial prior for each class, that is used to regularize the model. Although this approach yields results comparable to representation matching, using a spatial prior seems not reasonable when classifying aerial images, because objects can be located anywhere in the images. Class-balancing is also realized in the present work, however in a different way, because no explicit sample selection is used. An alternative approach based on semi-labelled samples is presented in (Iqbal and Ali, 2019). The authors propose to use spatially independent samples with a high confidence-score from the semi-labelled images acquired by aggregating predictions at multiple scales. Since their sampling

strategy relies on the assumption that the relative class-distributions in each image are similar in source and target domains this approach is probably not applicable to remote sensing applications, where large regions may contain only a single class. An approach for implicit instance transfer is presented in (Vu et al., 2019). Here, the entropy of the class predictions for each pixel is minimized which corresponds to increasing the probability of the most probable class. Thus, it is conceptually like the supervised training on semi-labelled samples. Besides the direct entropy minimization, the authors propose an adversarial approach that aligns the entropy distribution of source and target domain samples using a discriminator network. In both cases the model is regularized w.r.t. the predicted target domain class distribution, assuming it is close to the class distribution of the source domain. Again, limited to street scene segmentation, they show that the second approach slightly outperforms the direct entropy minimization, while both methods achieve results comparable to those based on representation transfer. The method in the present paper is also based on entropy minimization, but only the direct version is explored and extended by a pixel-wise weighting strategy. Further, while Vu et al. (2019) and Zou et al. (2018) jointly train on labelled source domain data and unlabelled target domain data, in this work the source training and the domain adaptation are done in two individual steps. This is assumed to be more practical when adapting a model from one source to several target domains, because source training must be carried out only once. Further, any pre-trained model could be adapted to a new domain without requiring any source domain samples for the actual adaptation. Both, Vu et al. (2019) and Zou et al. (2018) rely on prior assumptions regarding the class distribution of the target domain. While Vu et al. (2019) consider a distribution prior, assuming that class distributions of the source and target domains are similar, Zou et al. (2018) go one step further and use a spatial prior for each class based on the distribution of source domain samples. Both assumptions are not generally valid in remote sensing scenarios, due to the previously mentioned reasons. Consequently, the proposed method does not rely on any assumptions regarding the target domain class distribution. The method can be seen as a combination of entropy minimization to realize instance transfer and class-balancing to address imbalanced class distributions. However, the balancing is realized here by weighting each pixels' loss depending on its semi-label. Further, predicted object boundaries are excluded during the adaptation in order to improve the adaptation stability, which is not done in any of the mentioned publications.

3. METHODOLOGY

In this section, the proposed strategy for the supervised source training and the unsupervised adaptation of a FCN is presented. To that end, a formal description of DA according to Tuia et al. (2016) is given. In DA, a source domain D^S and a target domain D^T are considered, both associated with remotely sensed imagery. The domains are further associated with the joint distributions $P^S(X,C)$ and $P^T(X,C)$ of the image features X and the class labels C. In this paper, the setting of a homogeneous DA (Wang & Deng, 2018) is addressed, where the class structures C and the feature space X are assumed to be identical for both domains. The basic assumption of DA is that the joint distributions $P^S(X,C)$ and $P^T(X,C)$ are different, but related. The difference may be due to the marginal distributions of the features, i.e. $P^S(X) \neq P^T(X)$, or the posteriors, i.e. $P^S(C|X) \neq P^T(C|X)$. In both cases, the differences must not be too large. In the semi-supervised setting, a training data set T^S of labelled training samples is available in the source domain, each consisting of a tuple $(\mathbf{x}_i^S, \mathbf{c}_i^S)$ with $\mathbf{x}_i^S \in X$ and $\mathbf{c}_i^S \in C$ (in the addressed application, $(\mathbf{x}_i^S, \mathbf{c}_i^S)$ corresponds to a labelled image patch, hence $\mathbf{c}_i^S$ is a matrix with one class label per pixel in $\mathbf{x}_i^S$). The information available in D^T is restricted to the set U^T of unlabelled samples $\mathbf{x}_i^T \in X$. The task of SSDA is to use labelled data T^S and the unlabelled data U^T to learn a classifier that predicts the unknown labels $\mathbf{c}_i^T$ in the target domain.

In the proposed method this task is tackled by a two-step strategy. Firstly, a FCN is trained in a supervised way on labelled source domain training data T^S, resulting in the model M^S. The model is trained by minimizing the deviations between predicted labels and the reference $\mathbf{c}_i^S$, measured by a differentiable loss function $\mathcal{L}(M^S, \mathbf{x}_i^S, \mathbf{c}_i^S)$ as described in section 3.2. In the second step the model is adapted to a target domain D^T based on the unlabelled data U^T, resulting in the final target domain classifier M^T. The corresponding strategy is described in section 3.3.

3.1 Network Architecture

The FCN used in this work is designed to have a large receptive field while being able to propagate low level details through the network to preserve precise object boundaries. Preliminary experiments using different FCN architectures have shown that these two properties mainly affect the performance in AIC. They are commonly achieved by using encoder-decoder networks with skip-connections such as U-Net (Ronneberger et al., 2015). However, this architecture uses strong spatial down-sampling and deep feature maps, which leads to a large number of learnable parameters. For instance, the conventional U-Net has $\sim$30M parameters. Ronneberger et al. (2015) state, that convolutions with zero-padding should be avoided because they produce artefacts wherever the receptive field exceeds the boundaries of the input image. This is even more important when a larger receptive field is used. Nevertheless, zero-padding is frequently used in AIC applications, possibly resulting in larger training times, boundary artefacts or even a suboptimal performance.

The proposed architecture combines several techniques to enable a large receptive field, yet preserving low-level information without any of the listed drawbacks. This is mainly achieved by combining partial convolution based padding (Liu et al., 2018) with dilated convolutions (Fisher and Vladlen, 2016). While dilated convolutions can effectively increase the receptive field without heavily increasing the number of parameters, partial convolutions reweight the parameters of each learned convolutional filter in areas where padding is necessary, i.e. at the border of the input. The two concepts are combined in residual blocks. In each residual block, the input is convolved with four dilated convolutional layers with dilation rates of 1, 2, 3 and 4, respectively, using partial convolution based padding. The results are concatenated, merged by another convolutional layer and added to the input of the block. Using multiple filters with different dilation rates is inspired by the inception ResNet (Szegedy et al., 2017), where it was shown to improve the performance.

The architecture takes input patches of size 256×256 px containing both MSI and the rasterized height data. Firstly, a down-sampling layer is applied that performs a strided convolution (Springenberg et al., 2015) with a step with of 4 along both spatial dimensions. Next, 8 residual blocks are concatenated, followed by an up-sampling layer that uses a strided transposed convolution (Noh et al., 2015) again with a step width of 4 to scale the feature space back up to the size of the input. Down-sampling and up-sampling layer also use partial convolution based padding. The last layer predicts the class probabilities for each pixel using the softmax function. As activation function leaky rectified linear units (leaky ReLU) with a slope of 0.1 are used. All residual blocks as well as the up-sampling layer include a

dropout layer (Srivastava et al., 2014) to regularize the parameters and prevent overfitting to the training data. The architecture is presented in figure 1. All architecture-related parameters were found empirically in preliminary experiments.

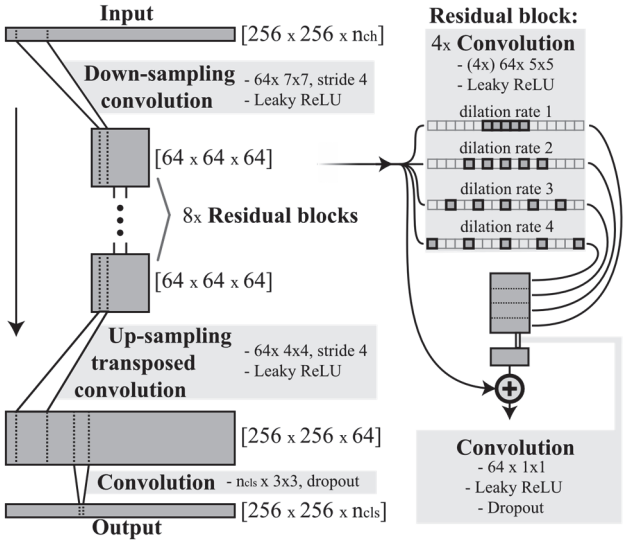

Figure 1: Proposed FCN architecture

In a typical scenario with $n_{ch} = 4$ input channels and $n_{cls} = 5$ classes, the model has about 3.5M parameters, one order of magnitude less than frequently used networks for image classification like U-Net. However, the dilated convolutions in each residual block result in a very large receptive field, in this configuration a theoretical window of over 512×512 px. Since the size of the receptive field is more than two times the input size, all predictions are affected by all input pixels. For instance, the prediction of the class label for the pixel in the lower-left corner is affected by the values of the pixel in the upper-right corner.

3.2 Supervised training

The supervised training of the model is based on minimizing $\mathcal{L}$ based on mini-batch gradient-descent. Instead of the commonly used cross-entropy loss $\mathcal{L}_{ce}$, the multi-class focal loss $\mathcal{L}_{fcl}$ is used here, because it is less affected by an unbalanced class distribution, often the case in aerial scenes. The focal loss was proposed by Lin et al. (2015) for binary image classification and adapted to the multi-class case in (Yang et al., 2019). Supervised training is carried out for a fixed number of iterations using the Adam optimizer (Kingma and Ba, 2015) and a constant learning rate. The batch size is not fixed but increased during training, which speeds up the training time without decreasing the classifiers performance. The tuning strategy and choice of all training-related hyper-parameters is further described in section 5.2.

During training, data augmentation is used to increase the variety of training samples and, thus, to improve the model generalization capability (Shorten and Khoshgoftaar, 2019). In AIC often only weak data augmentation is used, e.g. by random cropping of patches, followed by a random rotation in steps of 90° and a horizontal or vertical flip with a respective probability of 50% (Tasar et al., 2019). Sang and Minh (2018) perform only random flipping and Nogueira et al. (2019) do not use data augmentation at all. Yang et al. (2019) perform a slightly stronger geometrical augmentation, by rotating the patches in steps of 90°. In this paper a stronger data augmentation is applied, assuming that it will increase the model performance on the target domain. In each iteration a random affine transformation is applied in combination with radiometric augmentation, where each channel of the input patch is modified by a random linear transformation. This corresponds to a modification of brightness and contrast of each channel independently. Although the transformation of height data cannot be considered as radiometric augmentation, it is treated equivalently here assuming that a random transformation of the height data can compensate for differences in object heights or the ground level in D^S and D^T.

3.3 DA using entropy minimization

In this work, the actual DA step is carried out after source training as presented in section 3.2. The proposed approach realizes the concept of instance transfer in order to adapt the initial model M^S to a target domain D^T in an unsupervised way. The idea of instance transfer is to use semi-labelled samples from D^T, obtained by applying M^S to U^T to retrain the classifier on samples with high confidence, i.e. measured by the entropy of the prediction. In this work, the instance transfer is not realized by supervised training using semi-labelled samples, but instead, following (Vu et al., 2019), by directly minimizing the entropy E of predicted class distributions of target domain samples. Conceptually, this minimization is directly related to supervised training on semi-labelled samples, because minimizing the entropy corresponds to a maximization of the confidence of the currently most probable class. However, instead of explicitly choosing semi-labels with high confidence, this is done implicitly in entropy minimization. Samples with high confidence, thus a low entropy, result in larger gradients compared to samples with a high entropy and, thus, contribute more to the stochastic gradient descent. The entropy loss $\mathcal{L}_{ent}$ is derived as follows. Let n_{cls} be the number of classes, the entropy $E_{i,x,y}$ of the predicted class distribution $p_{i,x,y}$ for the pixel at position (x, y) in image i is defined as

$$E_{i,x,y} = -\log(n_{cls})^{-1} \cdot \sum_{c=1}^{n_{cls}} p_{i,x,y,c} \cdot \log(p_{i,x,y,c}), \qquad (1)$$

$\mathcal{L}_{ent}$ for a mini-batch with b_{ada} patches of size $h \times w$ becomes

$$\mathcal{L}_{ent} = \frac{1}{b_{ada} \cdot h \cdot w} \cdot \sum_{i=1}^{b_{ada}} \sum_{x=1}^{h} \sum_{y=1}^{w} E_{i,x,y} \ . \qquad (2)$$

Minimizing eq. 2 to perform the adaptation is assumed to be not reasonable when dealing with unbalanced class distributions, because the model would tend to increase the probability of the most frequent classes, thus, getting biased towards them. To counteract this behaviour a pixel-wise weighting strategy is proposed as follows. In each iteration of the adaptation phase the current model M is used to predict the semi-label map $\hat{\mathbf{c}}_i = argmax(M(\mathbf{x}_i))$ for each image $\mathbf{x}_i$ in the mini-batch. The weighting function $\Pi(c)$ for each class c corresponds to the ℓ_1-normalized, inverse class ratio, thus

$$\Pi(c) = \frac{o_c^{-1}}{\sum_{c'=1}^{n_{cls}} o_{c'}^{-1}} \ , \qquad (3)$$

where o_c is the number of pixels with semi-label c in all samples of the current mini-batch. As a second extension, pixel entropies that are closer to a predicted object boundary than ν pixels are not considered in the entropy minimization. This is motivated by the observation that classification models predict object boundaries usually with lower confidence, which leads to a high entropy of the corresponding pixels. Forcing a model to predict boundary regions with high confidence is assumed to be harmful during the adaptation. Formally, a binary boundary region indicator ϕ is introduced. Let B be the set of all pixels in the current mini-batch that have different semi-labels than any of its four adjacent pixels, the boundary region indicator becomes $\phi_{i,x,y} = 0$ if any

pixel *(i', x', y')* $\in B$ fulfils $\sqrt{(x-x')^2+(y-y')^2} \leq \nu$ and $\phi_{i,x,y} = 1$ otherwise. According to the predicted semi-label, the class weights and the boundary indicators, the weight for pixel *(i, x, y)* becomes $\gamma_{i,x,y} = \Pi_c(\hat{c}_{i,x,y}) \cdot \phi_{i,x,y}$. This leads to the proposed weighted entropy loss

$$\mathcal{L}^*_{\text{ent}} = \frac{1}{\Gamma} \cdot \sum_{i=1}^{b} \sum_{x=1}^{h} \sum_{y=1}^{w} \gamma_{i,x,y} \cdot E_{i,x,y}\,, \qquad (4)$$

where Γ denotes the sum of all weighting factors over the current mini-batch used to normalize the magnitude of the overall loss. During adaptation, $\mathcal{L}^*_{\text{ent}}$ is minimized using data samples from the training and test sets of the target domain. The adaptation is carried out for a fixed number of epochs using Adam optimizer for stochastic gradient descent. The selection of all hyper-parameters of the adaptation phase is described in section 4.2.

4. EXPERIMENTS

4.1 Datasets

For the evaluation of the proposed strategy datasets from 5 different German cities are used. Data from a sixth city was used for tuning purposes only. Firstly, the datasets *Potsdam (P)* and *Vaihingen (V)* are used, provided by the ISPRS labelling challenge (Wegner et al., 2017). Secondly, datasets for *Schleswig (S), Hameln (H), Buxtehude (B)* and *Nienburg (N)* are used. In the following, each city is treated as a separate domain. The datasets were pre-processed to have a common ground sampling distance of 20 cm, which required a spatial down-sampling of the data for *V* and *P*. Furthermore, all reference data were mapped to a common class structure, containing the five classes *Sealed ground*, *Building*, *Natural ground*, *Vegetation* and *Vehicle*. To this end, the class *Clutter* of the datasets *P, V, S* and *H* was manually re-labelled to one of the remaining classes. The classes *Water* and *Soil*, originally present in the reference for *S* and *H*, were mapped to *Natural ground*. The reference for *B* and *N*, provided by (Vogt et al., 2018), was revised to match the shared class structure. For all datasets the channels *near infrared* (NIR), *red* and *green* are available, thus these channels are used to compose the MSI. Normalized digital surface models (nDSM) are used as height maps the. While for *P* and *V* the original split into training and testing patches was kept, the remaining datasets were randomly image-wise split into disjoint sets for training and testing with a ratio of approximately 2:1, such that the class distribution of each subset roughly corresponds to the overall class distribution of the dataset. The dataset *N* was only used for tuning the hyper-parameters of the method. To obtain unbiased results, this dataset is not used for the evaluation of the method. Table 1 shows the size of training- and test set and the class distribution of each domain.

Domain		*P*	*V*	*B*	*S*	*H*	*(N)*
Train. set [px]		54M	12.5M	66.7M	17M	25M	66.7M
Test set [px]		31.5M	14.4M	33.3M	9M	12M	33.3M
Class distribution [%]	*Seal. Gr.*	31.1	28.2	22.1	14.1	18.8	22.3
	Building	25.7	26.0	19.7	14.7	19.1	18.4
	Nat. Gr.	25.9	21.6	36.9	38.9	36.2	42.1
	Vegetat.	15.5	22.9	20.3	31.5	24.5	16.6
	Vehicle	1.8	1.3	1.0	0.8	1.3	0.7

Table 1: Dataset overview.

In a pre-processing step the colour channels of all domains were normalized individually to zero-mean and unit-standard deviation based on the statistics of each domain. The new value $v'_{j,b,d}$ of band b of pixel j in domain d after normalization thus is computed as $v'_{j,b,d} = (v_{i,b,d} - \mu_{b,d})/\sigma_{b,d}$, where $\mu_{b,d}$ is the mean and $\sigma_{b,d}$ the standard deviation of all pixels of band b in domain d. The nDSMs were normalized according to $h'_{i,b,d} = h_{i,b,d}/u_h$ with a fixed value of $u_h = 5$ m to bring them to a value range close to the normalized MSI channels. For the evaluation of the proposed network architecture on the ISPRS labelling challenge, the original version of the *Vaihingen* dataset with a GSD of 8 cm was used, considering the original classes including *Clutter*. The channel-wise normalization was carried out as described above. The nDSM for *V* was provided by Gerke (2015).

4.2 Test setup and evaluation protocol

The evaluation is split into three parts. In the first experiment, the proposed network architecture is evaluated on the *Vaihingen* benchmark. The other two experiments correspond to the two phases of the proposed strategy for SSDA, i.e. source training and DA. To investigate the influence of data augmentation, two variants $\Omega \in [\Omega_-, \Omega_+]$ are used during source training. Variant Ω_- refers to a weak amount of augmentation. Samples, drawn from a random position, are randomly rotated by $n \cdot 90°$ with $n \in [0, 1, 2, 3]$ before being flipped horizontally or vertically with a respective probability of 50%. This variant is considered as frequently used augmentation strategy in AIC (see section 3.2). The second variant Ω_+ is the proposed, strong data augmentation strategy. Here, a random affine transformation is applied to obtain image patches using bilinear interpolation for input data and nearest neighbour interpolation for the label maps. The rotation is drawn from the uniform distribution $\mathcal{U}(0°, 360°)$, shear according to the normal distribution $\mathcal{N}(0, 0.3)$ and the scales for both spatial dimensions from $\mathcal{N}(1, 0.3)$. Additionally, each channel of the patch (including the height map) is linearly transformed with random bias r_b and scale r_c. Formally, the *j*-th channel of the sample $\mathbf{x}_j$ is transformed as $\mathbf{x}'_{i,j} = r_{c,j} \cdot (\mathbf{x}_{i,j} + r_{b,j})$. The corresponding random variables are drawn from $r_{c,j} \sim \mathcal{N}(1, 0.3)$ and $r_{b,j} \sim \mathcal{N}(0, 0.3)$. The resulting classifiers for each source domain D^S and augmentation scenario are denoted as $M^{S,\Omega}$. The proposed network architecture and all training related hyper-parameters were tuned in advance on domain *N*, such that the average overall accuracy on the test set of *N* based on both augmentation variants is maximised. In the tuning process, the following hyper-parameters were obtained. Training is done for 100K iterations using the Adam optimizer with a fixed learning rate of 10^{-4} and hyper-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch-size is initialised with 2 and increased by 1 every 6K iterations up to a maximum size of 16. The hyper-parameters related to the adaptation were obtained by maximising the average improvement of adapting the models from *N* to *H*. DA is done for 200 iterations by minimizing $\mathcal{L}^*_{\text{ent}}$ for mini-batches of target domain samples, obtained without data augmentation. The batch-size is set to 24, and the boundary margin to $\nu = 2$ px. The Adam optimizer is used with a learning rate of 10^{-6} and hyper-parameters $\beta_1 = 0.0$ and $\beta_2 = 0.99$.

All evaluations are done on the test set of the respective target domain. A sliding window evaluation is performed, i.e. the input image is split into overlapping patches (with an overlap of 50% in both spatial dimensions), each of the patches is processed and the resulting probability distributions for pixels in overlapping areas are averaged. Following Kaiser et al. (2017), the quality metrics *Overall Accuracy* (OA) and the *Mean F1-score* (MF_1) are used to assess the resulting label predictions. While the first one, being the ratio of correct predictions to the total number of predictions, can be biased for imbalanced class distributions, the MF_1 averages the prediction quality, i.e. the harmonic mean of precision and recall, of each class equally and, thus, is not biased towards classes with higher frequency.

4.3 Evaluation of the FCN architecture

In this section, the proposed network architecture is evaluated on the original *Vaihingen* dataset from the ISPRS labelling challenge. The model is trained using the protocol described in section 4.2 using augmentation scenario Ω_+. To validate the effect of using the focal loss, a second model is trained using the standard cross-entropy loss $\mathcal{L}_{ce}$. Two inference protocols (IP) are evaluated. The first (IP1) is the sliding window inference presented in section 4.2. The second protocol (IP2) additionally evaluates vertically and horizontally flipped versions of each image and averages the predictions of corresponding pixels, resulting in a higher redundancy per pixel. Table 2 shows the achieved quality measures, corresponding to those listed on the benchmark website (Wegner et al., 2017).

Loss	IP	Overall Acc. [%]	F_1-score [%]				
			Imp. sur.	*Build.*	*Low veg.*	*Tree*	*Car*
$\mathcal{L}_{fcl}$	1	90.5	92.2	95.7	83.8	89.5	83.2
	2	90.7	92.2	95.6	84.6	89.9	82.8
$\mathcal{L}_{ce}$	1	90.6	92.3	95.6	84.1	89.7	79.2
	2	90.8	92.6	95.7	84.5	89.9	80.1

Table 2. Achieved quality metrics for the *Vaihingen* benchmark.

Having followed the protocol of the ISPRS labelling challenge, i.e. evaluating the model on the reference with eroded class boundaries, the above results can be compared to those, achieved by other methods. Currently, the benchmark website lists the best OA as 91.6%. Thus, the achieved results are slightly worse (~1%). Regarding the loss function, for both IPs the OA is 0.2% higher when minimizing $\mathcal{L}_{ce}$, but The F_1-scores are less balanced. The F_1-score of the underrepresented class *Car* is around 3% higher, when minimizing $\mathcal{L}_{fcl}$ instead. Thus, it is reasonable to use $\mathcal{L}_{fcl}$ when dealing with unbalanced class distributions.

4.4 Evaluation of models before DA

In this experiment, models are trained in a supervised way using the labelled source domain training data of $D^S \in [P, V, B, S, H]$, and augmentation variant $\Omega \in [\Omega_-, \Omega_+]$. The resulting models are evaluated on the test set of all domains. Evaluating the models on the target domains without any adaptation technique is considered as baseline when assessing the effectiveness of DA in the further experiments. By training with two different augmentation variants, the initial statement, that strong data augmentation can partially alleviate the domain-gap, is validated. Table 3 shows the resulting OA and MF_1. Results printed in bold font correspond to intra-domain (ID) settings, thus $D^S = D^T$. The settings, where $D^S \neq D^T$ are referred to as cross-domain (CD) settings.

In nearly all cases, the quality metrics increase when changing the augmentation strategy from Ω_- to Ω_+. However, in the ID settings a stronger augmentation yields only a minor improvement of ~1% on average for both metrics, while the averaged OA for CD settings is increased by 8.4% and the average MF_1 by 10.5%. The worst overall results are obtained, when adapting from *S*, which may be due to the smallest amount of training samples in this domain. The best initial CD performance is obtained for *S*, when training on *B*, the domain with most available training samples. This is the only scenario, where training with weak augmentation outperforms the strong augmentation at least in OA. However, both variants outperform the supervised training on *S* with weak augmentation. After source training with Ω_+ the average MF_1 in the CD settings is 74.7%. Compared to the average MF_1 in the ID settings with 83.6% a gap of around 10% remains. Interestingly, this roughly corresponds to the remaining gap when training on very noisy labels (Kaiser et al., 2017).

D^S	D^T / Ω	*P* OA	*P* MF_1	*V* OA	*V* MF_1	*B* OA	*B* MF_1	*S* OA	*S* MF_1	*H* OA	*H* MF_1
P	–	**86.8**	**85.6**	73.6	62.3	72.3	61.0	64.0	53.2	75.4	62.6
	+	**87.8**	**86.6**	80.0	74.9	73.8	69.7	69.0	64.4	76.6	70.0
V	–	70.2	63.2	**86.2**	**81.4**	75.1	63.2	69.4	57.5	75.1	66.6
	+	76.6	72.5	**86.9**	**81.8**	80.6	74.9	73.8	68.8	74.4	67.2
B	–	66.0	63.0	72.0	66.9	**86.6**	**83.8**	85.6	79.2	73.0	65.5
	+	73.7	69.9	83.5	78.8	**87.6**	**84.3**	84.8	80.5	80.5	74.6
S	–	37.9	32.9	65.0	56.4	47.8	40.8	**84.1**	**77.5**	71.3	63.2
	+	64.3	57.5	79.3	71.2	76.9	72.2	**88.2**	**81.2**	71.5	71.4
H	–	64.5	58.6	76.4	69.6	69.6	77.9	77.9	68.7	**88.8**	**83.8**
	+	76.0	72.8	81.0	74.8	81.4	77.6	83.4	78.5	**88.9**	**83.9**

Table 3: Quality metrics in [%] of non-adapted models, trained on D^S and evaluated on the test set of D^T.

4.5 Evaluation of DA

The models obtained by source training with $\Omega \in [\Omega_-, \Omega_+]$ are now adapted to the other domains using the proposed DA strategy. Table 4 shows the achieved improvements of OA and MF_1 compared to the initial evaluation in table 3. Negative transfers are printed in bold font. Because *H* was used as target domain to tune the hyper-parameters of the adaptation method which is why the respective results should be taken with caution.

D^S	D^T / Ω	*P* OA	*P* MF_1	*V* OA	*V* MF_1	*B* OA	*B* MF_1	*S* OA	*S* MF_1	*(H)* OA	*(H)* MF_1
P	–	-	-	0.1	0.2	0.0	0.0	3.5	5.0	0.2	0.3
	+	-	-	1.6	1.4	4.1	4.6	2.5	4.3	1.2	4.2
V	–	2.8	5.5	-	-	0.6	5.1	0.8	0.8	1.4	0.8
	+	1.0	1.5	-	-	0.1	1.7	2.7	3.3	3.6	5.1
B	–	2.5	5.1	5.9	6.5	-	-	1.3	0.9	6.8	9.9
	+	3.3	3.9	0.6	0.3	-	-	0.8	0.9	4.5	6.3
S	–	10.1	11.1	6.5	6.4	3.4	4.6	-	-	0.2	0.7
	+	3.8	5.2	1.3	1.7	0.8	1.9	-	-	1.3	3.8
(H)	–	8.0	10.4	0.6	0.1	0.5	**-18.0**	2.7	5.3	-	-
	+	1.0	0.6	0.9	0.7	**-0.5**	0.2	0.3	0.1	-	-

Table 4: Improvements in [%], achieved by adapting $M^{S,\Omega}$ to D^T.

Adapting the models initially trained using strong augmentation leads to a positive transfer in 19 out of 20 cases with an average improvement of 1.8% in overall accuracy and 2.6% in MF_1. Although this is much smaller than the influence of data augmentation, in some cases a considerably large improvement of ~5% in MF_1 was achieved. After adaptation, the average CD metrics are 79.3% in OA and 74.7% in MF_1. Although adapting the models trained using weak data augmentation yields slightly higher improvements, the average CD metrics are still much lower with 72.0% in OA and 64.7% in MF_1. Interestingly, the adaptation from *H* to *B* leads to a negative transfer regardless of the augmentation strategy in source training, while adapting from *B* to *H* yields relative high improvements. It can be deduced that the performance of the adaptation is not symmetrical w.r.t. D^S and D^T. Figure 2 shows exemplary results, after adapting the models, trained with strong augmentation. The largest variations after adaptation can be observed in the class *Vegetation*, probably due to seasonal variations that highly affect the appearance of trees. This is clearly visible in the exemplary patch for *H*, captured in autumn, showing several trees without leaves. The images for the domains [*V, B, S*] were captured in summer, thus, do not contain any trees without leaves. Even after adaptation, the models barely detect any of the trees in *H*, while they are detected when coming from $D^S = P$, the only other domain captured in autumn.

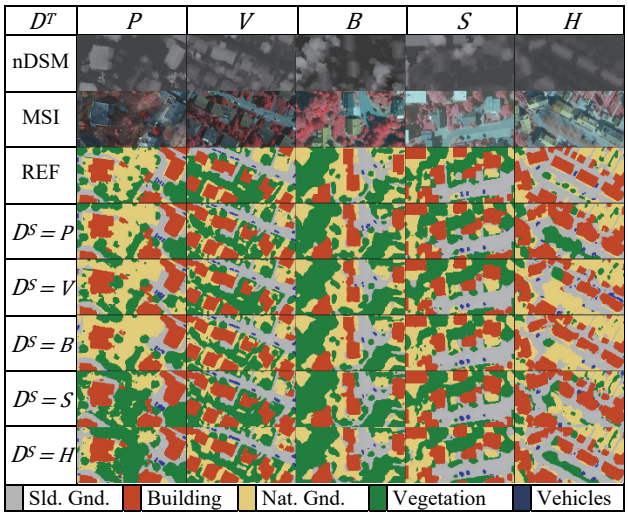

Figure 2: Exemplary test set samples and predictions. First three rows show nDSM, MSI and reference for a 256×500 region of each domain. Remaining rows show predictions after adaptation where $D^S \neq D^T$ and before adaptation where $D^S = D^T$.

4.6 Ablation study and comparison

In the last experiment, the influence of the proposed loss weighting strategy is investigated. To that end, the source models trained with Ω_+, are adapted by minimizing $\mathcal{L}_{ent}$, thus, the mean entropy without any weighting. Additionally, the direct entropy minimization strategy, as proposed in (Vu et al., 2019) is evaluated for comparison. Vu et al. (2019) propose a one-step approach, where the mixed loss $\mathcal{L}_{vu} = \mathcal{L}_{ce} + \lambda_{ent} * \mathcal{L}_{ent} + \mathcal{L}_{cp}$ is minimized. Here, $\mathcal{L}_{cp}$ is an additional loss that penalizes the deviation of source domain class distribution from target domain class distributions of each prediction; see (Vu et al., 2019) for further details. The relaxation parameter $\mathcal{L}_{cp}$ is set to $\mu = 0.5$ and the entropy weight to $\lambda_{ent} = 0.001$ as proposed by the authors. Although Vu et al. (2019) do not consider any online data augmentation, it is used here to have a fair comparison. Further, the training was started after the proposed source training. Training without online augmentation and training from scratch was carried out in additional experiments, not presented for lack of space. Both variants resulted in significantly worse results. Table 5 shows the resulting metrics, again as differences to the results obtained without adaptation. Negative transfers are printed in bold font.

	D^T	P		V		B		S		(H)	
D^S	$\mathcal{L}$	OA	MF_1	OA	MF_1	OA	MF_1	OA	MF_1	OA	MF_1
P	$\mathcal{L}_{ent}$	-	-	1.0	**-0.5**	2.5	0.5	2.3	0.4	**-3.3**	**-5.6**
	$\mathcal{L}_{vu}$	-	-	0.4	0.9	**-2.9**	**-2.7**	**-7.6**	**-5.7**	**-4.6**	**-2.7**
V	$\mathcal{L}_{ent}$	**-0.2**	**-3.2**	-	-	0.4	**-1.2**	1.6	0.3	0.2	**-1.6**
	$\mathcal{L}_{vu}$	**-4.4**	**-2.0**	-	-	**-0.2**	0.7	**-8.5**	**-6.3**	**-5.1**	**-4.7**
B	$\mathcal{L}_{ent}$	**-1.5**	**-4.2**	0.5	**-0.9**	-	-	1.4	0.5	**-2.1**	**-3.5**
	$\mathcal{L}_{vu}$	0.9	2.9	**-0.8**	**-0.6**	-	-	**-2.6**	**-2.2**	3.4	3.7
S	$\mathcal{L}_{ent}$	3.9	0.1	0.4	**-2.0**	0.7	**-1.6**	-	-	1.2	**-1.0**
	$\mathcal{L}_{vu}$	**-8.4**	**-6.8**	**-6.2**	**-5.0**	1.5	**-0.9**	-	-	**-2.5**	**-10.5**
(H)	$\mathcal{L}_{ent}$	**-0.1**	**-2.5**	1.1	**-1.4**	0.3	**-2.9**	1.6	**-0.3**	-	-
	$\mathcal{L}_{vu}$	**-5.4**	**-4.1**	**-0.8**	0.9	0.0	**-0.1**	**-1.6**	**-1.6**	-	-

Table 5: Improvements in [%], achieved by adapting $M^{S,\Omega+}$ to D^T using alternative loss functions.

The adaptation without the proposed class-balancing achieves a slight average improvement in OA of 0.1%, while the MF_1 decreases on average by 1.5%. A positive transfer w.r.t. both metrics could only be achieved in 5 out of 20 scenarios. The approach according to Vu et al. (2019) was even less stable with only 3 cases of positive transfer. The results indicate, that both variants perform worse than the proposed method in AIC scenarios with highly imbalanced classes. Adapting with the proposed strategy without excluding boundary regions and jointly training on labelled source domain samples was also carried out in non-listed experiments, both leading to a slightly worse performance.

5. CONCLUSION

In this paper, an approach for SSDA based on weighted entropy minimization was proposed and evaluated for several adaptation scenarios. The experiments indicate that a strong data augmentation can already alleviate the domain-gap significantly. In particular, the average MF_1 in in cross-domain settings was increased from 61.6% to 72.1%. By applying the proposed adaptation strategy, this metric was further increased to 74.7%. The adaptation approach is considered to yield mostly stable improvements, since only one out of 20 adaptation scenarios resulted in a negative transfer, regardless of the augmentation strategy during source training. In contrast, adaptation without the proposed weighting strategy resulted mainly in negative transfer, indicating that the proposed weighting strategy is necessary when dealing with imbalanced domains. The proposed FCN architecture performs comparable to the state of the art while having one order of magnitude fewer parameters than common architectures like U-Net. Despite the stability of the proposed method, the average cross-domain metrics after adaptation are still ~9% lower than the intra-domain metrics and seasonal effects were not fully compensated as shown in the visual evaluation. The results of this work also support the general assumption that training on larger datasets result in better generalizing models.

Future research should analyse whether the proposed method can be combined with other DA methods, e.g. based on image-to-image-translation (Tasar et al., 2019) or domain adversarial training (Wittich and Rottensteiner, 2019). Due to the obviously large impact of proper data augmentation during source training, it further seems reasonable to seek for more sophisticated augmentation scenarios that generate a wider range of meaningful augmentations or to deduce the augmentation parameters from statistical differences between the source- and target domain.

ACKNOWLEDGEMENTS

I thank the Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), the Landesamt für Vermessung und Geoinformation Schleswig Holstein and the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) for providing the datasets. Reference for the 3City dataset was provided by (Vogt et al., 2018).

REFERENCES

Benjdira, B., Bazi, Y., Koubaa, A., Ouni, K., 2019. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing 11(11),* 1369.

Bruzzone, L., Chi, M., Marconcini, M., 2006: A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing 44(11),* 3363-3373.

Fisher, Y., Vladlen, K., 2016: Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations (ICLR) 4.*

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Lorachelle, H., Laviolette, F., Lempitsky, V., 2016: Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(1), 2096-2030.

Gerke, M., 2015. Use of the stair vision library within the ISPRS 2d semantic labelling benchmark. Tech. rep., International Institute for Geo-Information Science & Earth Observation.

Hoffmann, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. *International Conference on Machine Learning (ICML)*, 1989-1998.

Huang, H., Huang, Q., Krähenbühl, P., 2018. Domain transfer through deep activation matching. *European Conference on Computer Vision (ECCV).* 590-605.

Iqbal, J. and Ali, M., 2019. MLSL: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling. *arXiv:1909.13776*.

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing 55 (11)*, 6054-6068.

Kingma, D. P., Ba, J. L., 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR).*

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems (NIPS'12) 25, Vol. 1*, 1097-1105.

Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *IEEE International Conference on Computer Vision (ICCV)*, 2980-2988.

Liu, G., Shih, K.J., Wang, T.C., Reda, F.A., Sapra, K., Yu, Z., Tao, A., Catanzaro, B., 2018. Partial convolution based padding. *arXiv:1811.11718*.

Long, J., Shelhamer, E., Darrell, T., 2015a. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431-3440.

Long, M., Cao, Y., Wang, J., Jordan, M. I., 2015b: Learning transferable features with deep adaptation networks. *International Conference on Machine Learning (ICML) 37,* 97-105.

Matasci, G., Volpi, M., Kanevski, M., Bruzzone, L., Tuia, D., 2015. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing 53(7),* 3550-3564.

Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. R., dos Santos, J. A. (2019). Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, *57(10)*, 7503-7520.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. *IEEE International Conference on Computer Vision (ICCV),* 1520-1528.

Pan, S. J., Yang, Q., 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22(10)*, 1345-1359.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 234-241.

Shorten, C., Khoshgoftaar, T. M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data 6(1),* 60.

Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: the all convolutional net. *International Conference on Learning Representations workshops.*

Sang, D. V., Minh, N. D. 2018. Fully Residual Convolutional Neural Networks for Aerial Image Segmentation. *Proceedings of the 9th International Symposium on Information and Communication Technology*, 289-296.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 1929-1958.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A., 2017: Inception-v4, inception-resnet and the impact of residual connections on learning. *31st AAAI Conference on Artificial Intelligence*, 4278-4284.

Tasar, O., Happy, S.L., Tarabalka, Y., Alliez, P., 2019. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *arXiv:1907.12859*.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: an overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine 4(2)*, 41-57.

Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2517-2526.

Vogt, K., Paul, A., Ostermann, J., Rottensteiner, F., Heipke, C., 2018. Unsupervised source selection for domain adaptation. *Photogrammetric Engineering & Remote Sens. 84(5)*, 249–261.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: a survey. *Neurocomputing 312*, 135-153.

Wegner et al., 2017. The ISPRS 2D semantic labelling contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (accessed 2/2/2020).

Wittich, D., Rottensteiner, F., 2019. Adversarial domain adaptation for the classification of aerial images and height data using convolutional neural networks. *Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7, 197–204.

Yang, C., Rottensteiner, F, Heipke, C., 2019. Towards better classification of land cover and land use based on convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XLII-2/W13, 139-146.

Zhang, Y., David, P. and Gong, B., 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. *IEEE International Conference on Computer Vision*, 2020-2030.

Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T., 2018. Fully convolutional adaptation networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6810-6818.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine 5(4)*, 8-36.

Zou, Y., Yu, Z., Kumar, B.V.K. and Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *European Conference on Computer Vision (ECCV),* 289-305.